Modelos Probabilistas Aplicados

Johanna Bolaños Zúñiga

Matricula: 1883900

Tarea 2

1. Libro Seleccionado – *The Scarlet Letter*

En el presente trabajo se realiza un análisis estadístico con datos directamente desde la Web, con el programa R versión 4.0.2 [3].

El análisis se llevará a cabo con el libro *The Scarlet Letter*, el cual se encuentra disponible de manera gratuita en *Project Gutenberg* [2].

2. Análisis estadístico

Para realizar el análisis estadístico y como ayuda visual se utilizaron diferentes tipos de gráficos. El código en R se encuentra en el repositorio de GitHub [1].

Para llevar a cabo el estudio, se visualiza el libro tanto por la cantidad de `letras` como de `palabras` que hay en él.

En la Figura 1, se muestran todos los caracteres Alfanuméricos utilizados en el libro. Sin embargo, nos interesan todos las letras, por lo cual, se procedió a realizar un filtro y, finalmente, obtenemos los datos mostrados en la Figura 2, en la cual se puede observar todas las letras y la cantidad de veces en que fueron empleadas.

Para visualizar de una mejor manera la información anterior, se organizan de manera decreciente la frecuencia de las letras, ver Figura 3. También, se elabora un histograma con 3 cubetas lineales, en la que podemos observar que son muy pocas las letras que tienen alta frecuencia de uso, ver Figura 4.

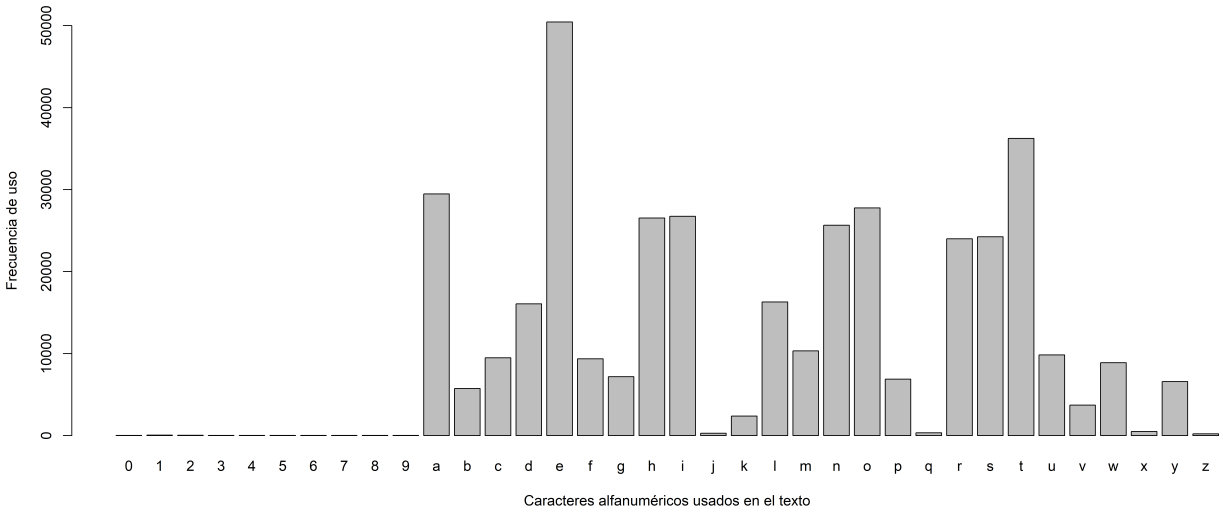

Figura 1: Frecuencia de los caracteres alfanuméricos del libro

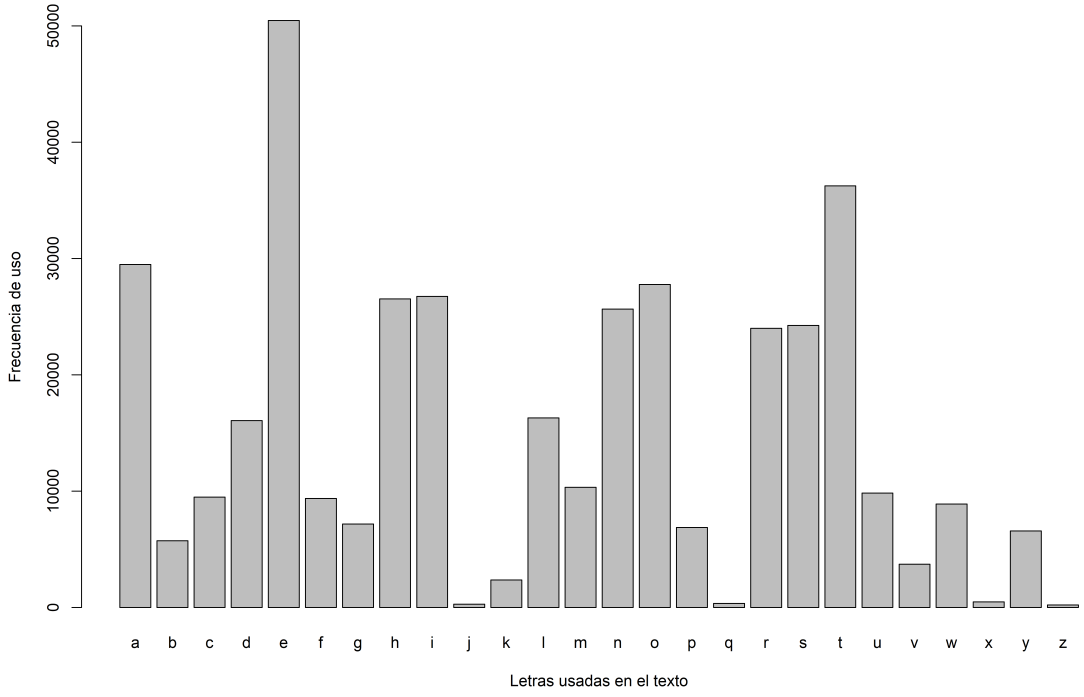

Figura 2: Frecuencia de las letras usadas en el libro.

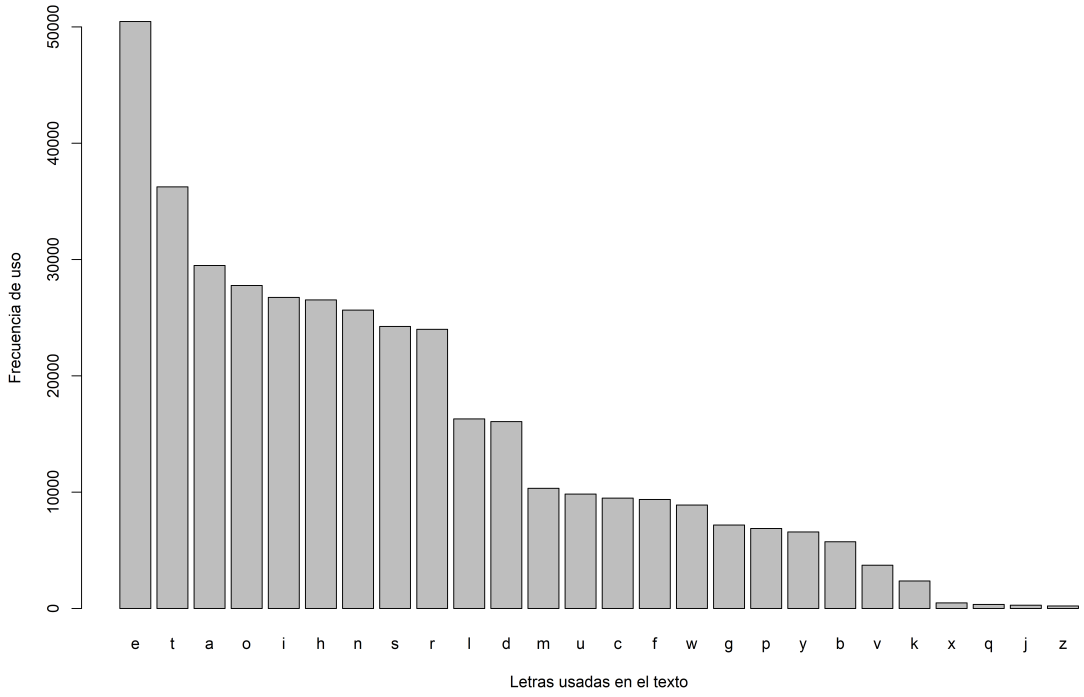

Figura 3: Frecuencia decreciente de las letras usadas en el libro.

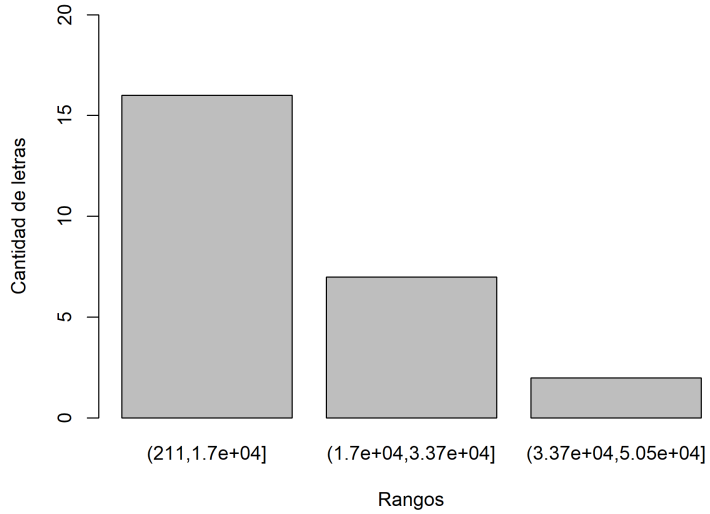

Figura 4: Histograma de la cantidad de letras usadas en el libro.

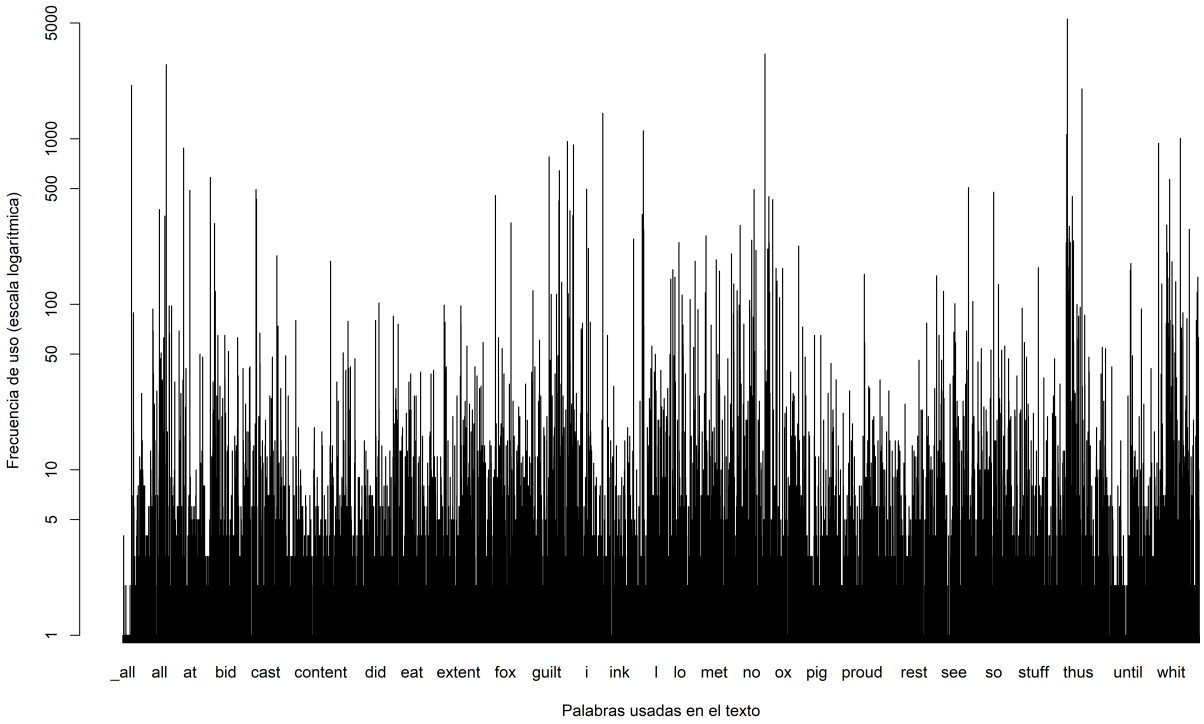

Figura 5: Palabras utilizadas en el libro

En cuanto al análisis por *palabras*, en la Figura 5 se muestran todas las palabras usadas en el libro. Sin embargo, se puede observar que son tantas las palabras utilizadas que es difícil hacer una interpretación a partir de ella. Por lo anterior, se realizó una depuración de las palabras que aportan poca información, como son las preposiciones, artículos y conjunciones. Estas palabras son conocidas como *stopwords*.

En la Figura 6, se muestran las palabras que tienen una frecuencia superior a 150. Para una mejor interpretación de los datos obtenidos, en la Figura 7 se organizan de manera decreciente las palabras más usadas.

De las palabras con mayor frecuencia podemos destacar que, la palabra más repetida es `Hester` y `Pearl`, lo cual tiene mucho sentido, ya que la primera es el nombre del personaje principal y la segunda, es el nombre de la hija. Sin embargo, a pesar de que `Prynne` es el apellido de la protagonista , este no es tan frecuente. En la Figura 8 se muestra de una manera más didáctica las palabras más usadas en el libro, donde el tamaño es mayor para las palabras que aparecen con más frecuencia, este tipo de gráfica se conoce como `nube de palabras` o en inglés *world cloud*.

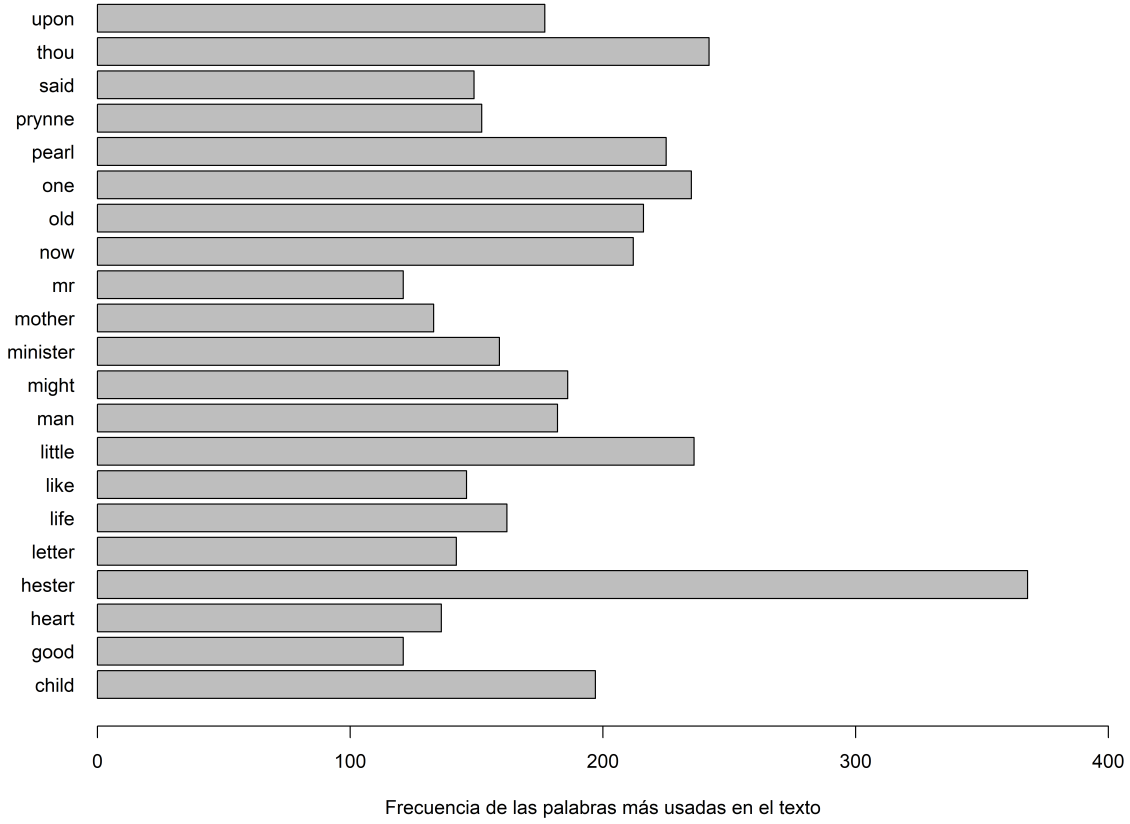

Figura 6: Palabras de mayor frecuencia en el libro

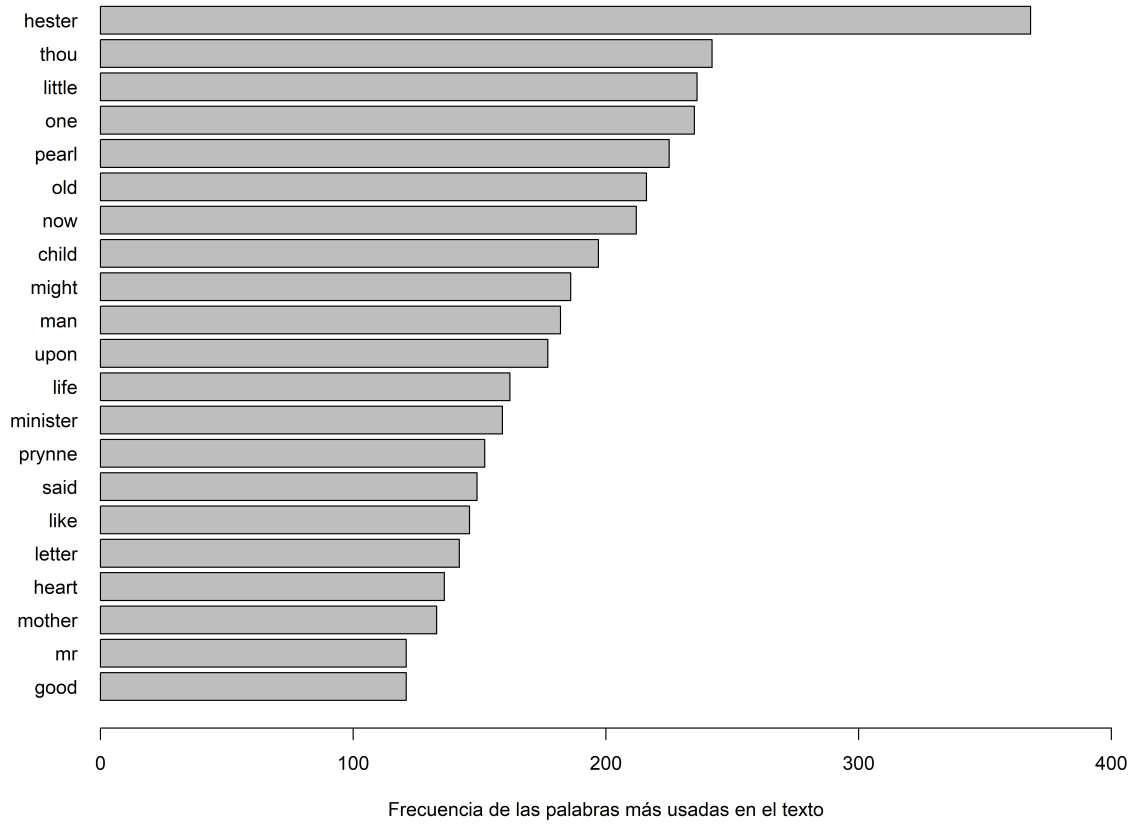

Figura 7: Frecuencia decreciente de las palabras más usadas en el libro.

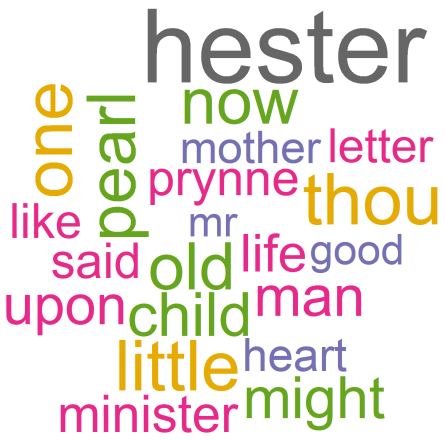

Figura 8: Nube de palabras más frecuentes en el libro

Referencias

[1] Bolaños Z., Johanna. Repositorio en GitHub de la clase de modelos probabilistas aplicados. Recursos libre, disponible en github.com/JohannaBZ/Probabilidad/tree/master/Tarea2, 2020.

[2] Hawthorne, Nathaniel. The Scarlet Letter. Recurso libre, disponible en http://www.gutenberg.org/ebooks/25344, 1985.

[3] The R Foundation. The R Project for Statistical Computing. https://www.r-project.org/, 2020.